HAL
open science

Xart system: discovering and extracting correlated arguments of n-ary relations from text

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie-Barthelemy, Mathieu Roche

To cite this version:

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie-Barthelemy, Mathieu Roche. Xart system: discovering and extracting correlated arguments of n-ary relations from text. 6th International Conference on Web Intelligence, Mining and Semantics (WIMS 2016), Jun 2016, Nimes, France. pp.8:1-8:12, 10.1145/2912845.2912855 . hal-01357738

HAL Id: hal-01357738

https://agroparistech.hal.science/hal-01357738v1

Submitted on 6 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Xart system: discovering and extracting correlated arguments of n-ary relations from text

Soumia Lilia Berrahou
Lirmm - 860 rue de St Priest, 34095 Montpellier, France
IATE - 2, place Pierre Viala, 34060 Montpellier, France
berrahou@lirmm.fr

Patrice Buche
IATE - 2, place Pierre Viala, 34060 Montpellier, France
buche@supagro.inra.fr

Juliette Dibie
UMR MIA-Paris - AgroParisTech/INRA, Université Paris-Saclay, 75005 Paris, France
dibie@agroparistech.fr

Mathieu Roche
CIRAD/IRSTEA/AgroParisTech - TETIS - 500, rue J.F. Breton, 34093 Montpellier, France
mathieu.roche@cirad.fr

ABSTRACT

In this paper, we present Xart system based on a hybrid method using datamining approaches and syntactic analysis to automatically discover and extract relevant data modeled as n-ary relations in plain text. A n-ary relation links a studied object with its features considered as several arguments. Our work focuses on extracting those arguments in text in order to populate a domain Ontological and Terminological Resource (OTR) with new instances.
Our approach relies on a new data representation in order to increase data expressiveness in the knowledge discovery process, using the concepts defined in the OTR.
Using sequential rules and pattern mining allows the discovery of implicit forms of expression that are used to describe arguments of n-ary relations in text. Once the implicit rules are discovered in specific patterns, we define as Ontological Sequential Patterns (OSP), we use syntactic relations to enrich the patterns in order to obtain Ontological Linguistic Sequential Patterns (OLSP) where the arguments of n-ary relations are expressed according to different levels of word abstraction (word, grammatical category and concept). We have made concluding experiments on a corpus from food packaging domain where relevant data to be extracted are experimental results on packagings. We have been able to extract up to 4 correlated arguments with a F-measure from 0.6 to 0.8.

Keywords

Information extraction; n-ary relation; argument; Ontological and Terminological Resource; datamining; sequential pattern; sequential rule; quantitative data; syntactic analysis; lingusitic pattern

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

1. INTRODUCTION

Discovering and extracting information reported in documents is a crucial stake in several domains in order to be able to reuse, manage, exploit, analyze the information they contain, and use them for decision making purpose. Our methodology enables the processing of documents on the Web, i.e. specialised databases like PubMed. Thus, our methodology helps to highlight knowledge discovered in textual documents in order to develop "Open Data Science" domain. Let us consider two examples of sentences (1) and (2) containing relevant information in two distinct domains: in food packaging domain and in the field of civil aviation. In sentence (1), a studied object (i.e. *polypropylene* film) is analyzed according to different features represented by quantitative data, associated with their numerical value and their unit (i.e. *thickness, oxygen permeability, temperature* and *relative humidity (RH)*). In sentence (2), the studied object is a plane *A380-800* and its features associated with their numerical value and their unit, are *transport capacity, flying range, speed*.

(1) Eight apple wedges were packaged into polypropylene trays and wrap-sealed using a 64 μm thickness polypropylene film with a permeability to oxygen of 110 $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ at 23 ° C and 0 % RH

(2) The A380-800 has a 150 tons of transport capacity, a 15 400 kilometers of flying range that allow a non-stop New York-Hong Kong flight with a 900 km/h up to 1012 km/h of speed

The relevant information extracted from these two sentences can be considered as instances of n-ary relations which can help domain experts in a decision making context. Nevertheless, instances of n-ary relations are painful to automatically identify and extract in text because arguments are rarely expressed in a unique sentence but in several sentences, usually in implicit and various forms of expression. Moreover, the expression of quantitative arguments

frequently varies in their attributes, i.e. the numerical value and the unit of measure, from a studied object to another one.
Those issues bring together several domains such as Natural Language Processing (NLP) and Knowledge Engineering. Indeed, the documents containing relevant information to be extracted use natural language combined with domain-specific terminology that is extremely tedious to extract in text. Information extraction is a particular domain of NLP which purpose is the extraction of specific knowledge such as named entities or relations. In NLP, a relation usually refers to a connection between entities in text.
In biomedical field, substantial amount of work on binary relation extraction is proposed. The first approaches to discover relations between entities focused on limited linguistic context and relied on discovering cooccurrences and pattern matching manually designed as in [15]. Rule-based techniques defined in forms of regular expressions over words or part of speech (POS) tags are used to construct linguistic or syntactic patterns [14], [23], [13], [24]. However, manually defined rules require heavy human effort.
Later, machine learning-based approaches, e.g. Support Vector Machines (SVMs) [20], were widely employed [26], [31], [21], [28] more precisely as a classification problem [25]. Those methods are successfully used but require a large amount of annotated data for training that usually takes tremendous human efforts to build, need many features and are based on numerical models not directly understandable by the final user.

The extraction task of n-ary relation is a more complex issue due to several arguments involved in the relation. The arguments can be expressed in implicit forms of expression and usually appear on several sentences. [19] conducted the first work on n-ary relation extraction expressed in a sentence and proposed, after identifying all binary relations between entities, to construct a graph of entities where edges denote binary relations. N-ary relation instances were finally constructed by finding the maximal cliques in the graph.
Later, other work were conducted dividing n-ary relation extraction issue on three main steps: the first step consists in identifying entities (or arguments) using resources such as ontologies or dictionaries, the second step consists in identifying the trigger word of the relation using dictionary-based methods or rule-based approaches to construct patterns from dependency parse results [18], or using machine learning methods [8], [7], [5], [32] in order to predict the trigger word of the relation. Finally in the third step, binary relations are constructed using the trigger word and machine learning methods are used to classify whether binary relations belong to the searched n-ary relation but with a substantial loss of accuracy. Relation extraction methods are all based on those three independant steps. In this work, we propose to consider another perspective on argument expression involved in n-ary relations. Instead of identifying data involved in n-ary relations, we are interested in discovering the implicit relationships between them occurring in text, using datamining approaches. Those techniques have already been succesfully used on text to discover implicit relations between entities in order to enrich ontologies [10] or to discover linguistic patterns without external resources [3], [9]. Moreover, in [16], authors use association rules and sequential patterns in order to propose comprehensive and reusable rules for text categorization.

In this paper, we present Xart system that relies on a domain OTR and takes advantages of datamining methods in order to discover implicit rules of expression between arguments involved in the searched relation. Then, the system uses syntactic analysis in order to construct Ontological Linguistic Sequential Patterns (OLSP) where the arguments of n-ary relations are expressed according to different levels of word abstraction (word, grammatical category and concept).
The paper is structured as follows. In section 2, we give a quick overview of Xart system. In section 3, we describe the step of Xart system based on our knowledge discovery process to discover Ontological Sequential rules and Patterns (OSP). In section 4, we describe the step of Xart system based on our hybrid approach to construct Ontological Linguistic Sequential Patterns (OLSP). In section 5, we firstly present and discuss the discovery of OSP and secondly, we present and discuss the results we obtained on the extraction of correlated arguments from text using our OLSP. In section 6, we finally conclude on our work.

2. XART SYSTEM

In this section, we present our Xart system to extract correlated arguments of n-ary relations in text. The step shown in figure 3 uses datamining approaches to discover specific sequential patterns, we defined as OSP of correlated arguments in text. The step shown in figure 4 proposes to enrich, extend discovered patterns with a specific syntactic information in order to construct OLSP and extract correlated argument instances of n-ary relations from text. In this section, we present the OTR, then propose the first definitions on which our work relies and finally we give a quick overview of Xart system.

OTR.

Relevant data are modeled as n-ary relation where a studied object is modeled as a symbolic argument and its features as quantitative arguments associated with their attributes, the numerical value and the unit of measure. Our representation of n-ary relations is the one of naRyQ (n-ary Relations between Quantitative data) [27]. naRyQ contains two components, a terminological component that gathers all terms of the domain, e.g. the names of packagings, and a conceptual component. The conceptual component of naRyQ is composed of a core ontology to represent n-ary relations and a domain ontology to represent specific concepts of a given application domain.

Figure 1 gives an excerpt of naRyQ in the food packaging domain on which rely our evaluations. In the up core ontology, generic concepts Relation_Concept and Argument represent respectively n-ary relations and their arguments. The n-ary relation concepts may be hierarchically organized by the specialization relation, shown by the lines in figure 1. In the down core ontology, generic concepts Dimension, UM_Concept, Unit_Concept and Quantity allow the management of quantities and their associated units of measure. The sub-concepts of the generic concept Symbolic_Concept represent the non numerical arguments of n-ary relations. The domain ontology contains specific concepts of a given application domain. They appear in naRyQ as sub-concepts of the generic concepts of the core ontology. The terminological component of naRyQ contains the set of terms de-

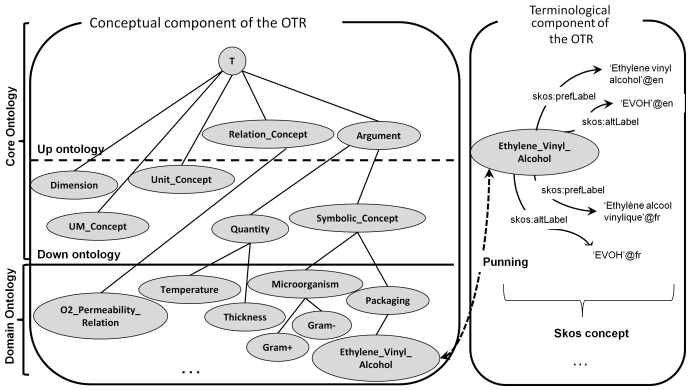

Figure 1: An excerpt of the concept hierarchy of naRyQ in food packaging domain

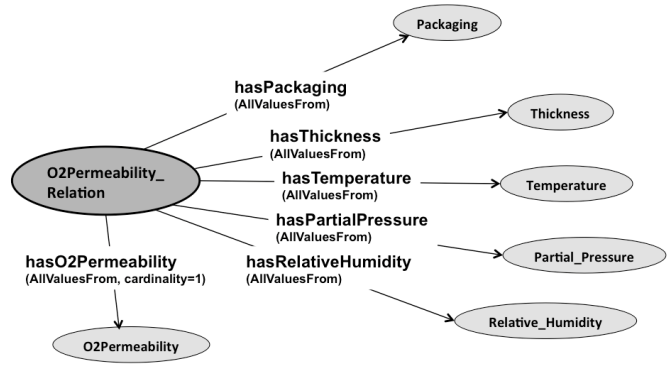

Figure 2: The n-ary relation O2Permeability_Relation

scribing the studied domain.

A n-ary relation is represented by a concept which is linked to its arguments by binary relations such that none of those arguments has a specific role (e.g. subject or object). We give a formal definition of such a representation of n-ary relations between quantitative data.

Definition 1.
Let us consider the ontology $OTR = \langle \mathcal{C}_{OTR}; \mathcal{R}; \mathcal{I}; \mathcal{V}; \leq_o \rangle$.
$\mathcal{C}_{OTR}$ *is a set of concepts in the ontology* OTR*;*
$\mathcal{R}$ *is a set of relations into* $\mathcal{C}_{OTR} \times \mathcal{C}_{OTR}$*;*
$\mathcal{I}$ *is a set of instances;*
$\mathcal{V}$ *is a set of values;*
$\leq_o$ *is a specialisation relation in* $(\mathcal{C}_{OTR} \times \mathcal{C}_{OTR}) \cup (\mathcal{R} \times \mathcal{R})$*;*

A **n-ary relation concept** $rel \in \mathcal{C}_{OTR}$, $\leq_o$*(rel, Relation_Concept) is defined in* OTR *by the set of binary relations* $r_j \in \mathcal{R}$ *which link the n-ary relation to its arguments, this set being composed of at least two binary relations:*

$$Def(rel) = \{r_j(rel, a_j) \mid r_j \in \mathcal{R}, \\ (a_j \in \mathcal{C}_{OTR} \wedge \leq_o (a_j, Argument))\}, \\ such\ that\ |Def(rel)| \geq 2$$

A n-ary relation is caracterized by its signature, i.e. the set of its arguments.

Definition 2.
Let us consider the ontology $OTR = \langle \mathcal{C}_{OTR}; \mathcal{R}; \mathcal{I}; \mathcal{V}; \leq_o \rangle$.
The **signature** *signatureR:* $\mathcal{C}_{OTR} \longrightarrow 2^{\mathcal{C}_{OTR}}$ **of a n-ary relation concept** $rel \in \mathcal{C}_{OTR}$ *as defined in Definition 1 is:*

$$signatureR(rel) = \\ \{(a_j \in \mathcal{C}_{OTR} \wedge \leq_o (a_j, Argument)) \mid \\ r_j(rel, a_j) \in Def(rel)\}$$

An example of n-ary relation concept is given in figure 2 and represents O2Permeability relation in naRyQ_pack OTR (food packaging domain OTR). The signature of the n-ary relation O2Permeability_Relation is:
signatureR(*O2Permeability_Relation*)= {Packaging, Thickness, Temperature, Partial_Pressure, Relative_Humidity, O2Permeability}.

Relevant feature.
In our work, we have defined units of measure associated with quantatitative arguments as relevant features in text to define an optimal context for discovering the searched arguments [4]. From this hypothesis, we propose two relevant textual contexts of search :

Definition 3. *(Pivot sentence)*
A pivot sentence is defined as the sentence where at least one unit referenced in the OTR is identified

Definition 4. *(Textual window)*
A textual window noted f_{sn} *is defined as the set of sentences composed of the pivot sentence and the n previous sentences and/or the n afterward sentences, where n is the window dimension. The search direction in sentences, noted s, is represented with - if we consider previous sentences, with + if we consider afterward sentences and* $\pm$ *if we consider previous and afterward sentences*

Those definitions are the basis of our work and allow us to define textual contexts in order to discover relevant information about n-ary relations over the steps of Xart system.

Discovering correlated arguments in ontological sequential patterns.
Identified units from the OTR are used to define textual contexts that allow us to stay close to the arguments of n-ary relations in text. As we have previously defined, relevant textual contexts are defined from units as textual windows. Textual windows are relevant set of sentences to discover implicit rules and patterns using datamining approaches. In this step, we use a new data representation that relies on the conceptual level defined in the OTR. This new data representation allows to increase data expressiveness in text and discover new patterns in the knowledge discovery process shown in figure 3. Those new patterns are defined as OSP since they contain conceptual level of argument expression.

Using syntactic analysis.
Finally, in this step shown in figure 4, we use specific syntactic relations, we defined as OSR as we will detail in section 4, in order to construct ontological linguistic sequential patterns.

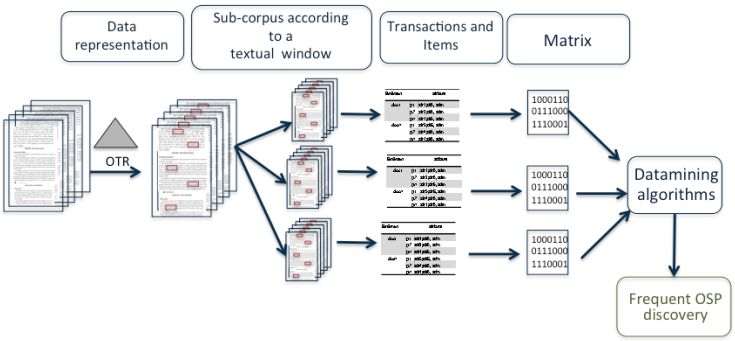

Figure 3: Knowledge Discovery Process of Xart system (KDP)

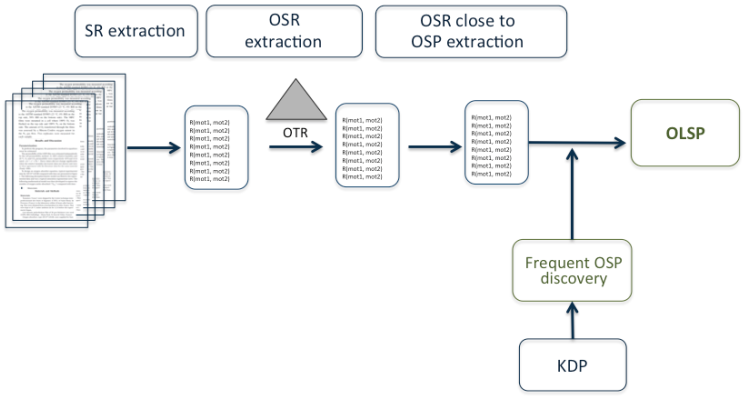

Figure 4: The hybrid approach of Xart system

3. DISCOVERING CORRELATED ARGUMENTS

In order to discover frequent patterns involving arguments of n-ary relations in plain text, we focus on the question about the existence of implicit relations in the expression of those arguments that could facilitate their extraction and their linking in the searched instance. For that purpose, using datamining approaches present significant advantages in discovering not only implicit rules from data but also sequential patterns in order to better understand how arguments are associated in text. In this section, we firstly present the basic definitions of datamining used in the paper, we secondly highlight our data representation, which relies on the OTR, in order to apply efficiently the knowledge discovery process on plain text, and we finally detail our knowledge discovery process to discover frequent patterns of correlated arguments.

3.1 Preliminaries

The following definitions are detailed in [2] and presented here according to an example taken from [11]. The database $\mathcal{DB}$ in table 1 represents a set of transactions. Each transaction represents the set of events (items) appearing in each city at several months. For each city, a sequence is generated as shown in table 2 and is composed of itemsets where an item can appear several times in a same sequence.
In section 3.3, we show how we structure and represent a text in order to select the transactions and items to mine. Each transaction corresponds to a set of relevant sentences containing arguments of n-ary relations and selected items are specific words in the neighborhood of identified arguments.

Definition 5. *(Sequence)*
Let $\mathcal{I} = \{I_1, I_2, ..., I_m\}$ be a set of items. An itemset is a non empty, non ordered set of items noted $(I_1, I_2, ..., I_k)$ where $I_j \in \mathcal{I}$. A sequence is a non empty ordered list of itemsets noted $< IS_1 IS_2 ... IS_p >$ where $IS_j \in \mathcal{IS}$, with $\mathcal{IS}$ the set of itemsets.

Extracting sequential patterns consists in searching frequent sub-sequences from sequences.

Definition 6. *(Sub-sequence)*
A sequence $A = < IS_1 IS_2 ... IS_p >$ is a sub-sequence of an other sequence $B = < IS'_1 IS'_2 ... IS'_m >$ $(A \preceq B)$ if $p \leq m$ and if there exists integers $j_1 < j_2 < ... < j_k < ... < j_p$ such as $IS_1 \subseteq IS_{j_1}, IS_2 \subseteq IS_{j_2}, ..., IS_p \subseteq IS_{j_p}$.

Example 1.
Consider the sequences in table 2, the sequence $S = < (Sun)(Heat = High) >$ is supported by sequences S_{Nimes} and $S_{Montpellier}$ as $S \preceq S_{Nimes}$ and $S \preceq S_{Montpellier}$.

A sequential pattern is a frequent sub-sequence characterized by a support, which represents the number of occurrences of a pattern in $\mathcal{S}$, a set of sequences. Extracting frequent sequential patterns is extracting patterns with a support value greater than a minimum support parameter θ . Let $\mathcal{M}$ be a set of extracted sequential patterns, then $\forall M \in \mathcal{M}$, $Support(M) \geq \theta$.

Definition 7. *(Sequential pattern support)*
A sequence $S \in \mathcal{S}$ supports a sequential pattern M when $M \preceq S$. The support of M is the number of sequences in S in which M is included. Let $\mathcal{S}'$ be the set of sequences that support M, then $\mathcal{S}' = S_i \in \mathcal{S}$ such as $M \preceq S_i$ and Support(M) = $|\mathcal{S}'|$.

Definition 8. *(Association rule)*
An association rule is an implication expression of the form $X \longrightarrow Y$ where X and Y are disjoint itemsets, i.e., $X \cap Y = \emptyset$. The strength of an association can be measured in terms of its support and confidence. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain X.

Example 2.
Consider the rule $\{Sun \longrightarrow Humidity = Low\}$. Since the support count for $\{Sun, Humidity=Low\}$ is 2 and the total number of transactions is 5, the rule's support is 2/5=0.4. The rule's confidence is obtained by dividing the support count for $\{Sun, Humidity=Low\}$ by the support count for $\{Sun\}$. Since there are 3 transactions that contain Sun, the confidence for the rule is 2/3=0.67.

The interpretation of a sequential rule is given by [12] as if the items of X occur in some transactions of a sequence, the items in Y will occur in some transactions afterward from the same sequence.

Now that the basic definitions used in this paper are given, we present our approach on data representation in order to optimize the expression of arguments in text and apply datamining algorithms.

3.2 Data representation

Stay close to the involved arguments.

As previously defined, we use units referenced in the OTR in order to define textual windows, which represent a set of relevant sentences close to the arguments of n-ary relations. Using textual windows, we get several sub-corporus. Each one is mined with datamining algorithms in order to discover sequential rules and patterns on the expression of arguments of n-ary relations. Both Sequential rules and patterns are useful to explore in order to identify interesting rules and relations on the expression of arguments in text, and sequential patterns are particularly useful for extraction pattern construction of arguments in text.

Let arguments express themselves.

The aim of this work is to extract argument instances, which forms of expression frequently change in text and which numerical values frequently change according to the measurements made on the studied object. Mining frequent patterns directly on text without increasing expressiveness of n-ary relation arguments substantially decreases knowledge discovery efficiency. We propose to tackle this issue by taking into consideration data expressiveness using a new representation. This new representation relies on the core ontology where symbolic argument and quantitative argument (i.e. quantity) are distinct. We propose with the definition 9 to increase the expressiveness of the symbolic arguments by representing them with their corresponding concepts, sub-concepts of the generic concept
$< Symbolic_Concept >$. For example, in our experiments

City	Month	items
Nîmes	01/2011	Humidity=low, Sun
Montpellier	02/2011	Sun
Nîmes	02/2011	Heat=High
Montpellier	03/2011	Humidity=Low, Heat=High, Sun
Nîmes	04/2011	Heat=Low, Wind

Table 1: Database $\mathcal{DB}$

City	Sequence
Nîmes	<(Sun Humidity=Low)(Heat=High)(Heat=Low Wind)>
Montpellier	<(Sun)(Humidity=Low, Heat=High, Sun)>

Table 2: Sequences $\mathcal{S}$ for each city

led on the corpus of packaging domain, we choose the sub-concept $< Packaging >$ according to the specialization relation (Packaging, Symbolic_Concept) in order to represent the studied object packaging in text.

Definition 9. *(Symbolic concept representation)*
$\mathcal{C}_{\mathcal{OTR}}$ is a set of concepts of the OTR;
$\mathcal{C}_{\mathcal{OTR}} = \mathcal{C}_{Rel} \cup \mathcal{C}_{Qty} \cup \mathcal{C}_{Symb}$ where $\mathcal{C}_{Rel}$ is a set of relations, $\mathcal{C}_{Qty}$ is a set of quantitative concepts, $\mathcal{C}_{Symb}$ is a set of symbolic concepts;
$\mathcal{W}_{\mathcal{OTR}}$ is a set of words in the terminological component of the OTR where all words $w_i \in \mathcal{W}_{\mathcal{OTR}}$ denote a concept $C_i \in \mathcal{C}_{\mathcal{OTR}}$;
$\forall$ t, term of the text, such that $\exists\ w_i \in \mathcal{W}_{\mathcal{OTR}}$ which denotes $C_i \in \mathcal{C}_{Symb}$, and $w_i = t$, and $\exists\ C_j$ such as $C_j \geq C_i$, and $C_j \in SignatureR(rel)$ with $rel \in \mathcal{C}_{Rel}$, then t is annotated by $C_j \in \mathcal{C}_{Symb}$ in the new data representation.

We propose with the definition 10 to increase the expressiveness of the quantitative arguments by increasing the expressiveness of their numerical values using their associated units of measure. Numerical values represent the relevant information we want to discover and which often vary, depending on the measurements made on the studied object. Units of measure related to numerical values are associated with specific sub-concepts of the generic concept $< Quantity >$. We use those sub-concepts of $< Quantity >$ to represent numerical values. We can therefore simply represent the quantity with the generic concept $< Quantity >$ and the unit with $< um >$ for $< Unit_Concept >$.

Definition 10. *(Quantity concept representation)*
$\mathcal{C}_{\mathcal{OTR}}$ is a set of concepts of the OTR;
$\mathcal{C}_{\mathcal{OTR}} = \mathcal{C}_{Rel} \cup \mathcal{C}_{Qty} \cup \mathcal{C}_{Symb}$ where $\mathcal{C}_{Rel}$ is a set of relations, $\mathcal{C}_{Qty}$ is a set of quantitative concepts, $\mathcal{C}_{Symb}$ is a set of symbolic concepts;
$\mathcal{V}$ is a set of values associated with the concepts $\mathcal{C}_{Qty}$ of the OTR ;
$\mathcal{I}_{UM} \subset \mathcal{I}$, subset of instances which represents units of measure;
$\mathcal{W}_{\mathcal{OTR}}$ is a set of words in the terminological component of the OTR where all words $w_i \in \mathcal{W}_{\mathcal{OTR}}$ denote a concept $C_i \in \mathcal{C}_{\mathcal{OTR}}$;
$\forall\ t_u$, term of unit in the text defined as $\exists\ w_i \in \mathcal{W}_{\mathcal{OTR}}$ such as w_i denotes $i \in \mathcal{I}_{UM}$ such that $i \in hasUnit(C_i)$ with $C_i \in \mathcal{C}_{Qty}$ and $C_i \in SignatureR(rel)$ with $rel \in \mathcal{C}_{Rel}$,
$\forall\ v_i$, a value in the text associated with t_u such as $v_i \in \mathcal{V}$, is annotated by $C_i \in \mathcal{C}_{Qty}$, C_i by <quantity> and t_u by <um> in the new data representation.

In example 3, the numerical value *64* is followed by the unit μm that is associated with $< Thickness >$ concept. Thus, $< numvalthick >$ is used to annotate each value followed by a unit associated with the concept $< Thickness >$, $< numvaltemp >$ is used to represent each value followed by a unit associated with the concept $< Temperature >$, and we simply represent the unit with $< um >$ for $< Unit_Concept >$. The data representation principle is shown in example 3: in sentence (1), the arguments that should express themselves in text are underlined. The sentence (2) corresponds to our new data representation of the sentence (1). This sentence contains an instantiation of the n-ary relation O2_Permeability Relation of figure 2, which represents the O2 permeability of a packaging in given experimental conditions, defined by the packaging thickness (64 μm), the temperature (23 ° C) and the relative humidity (0%).

Example 3.

(1) Eight apple wedges were packaged into polypropylene trays and wrap-sealed using a 64 μm thickness polypropylene film with a permeability to oxygen of 110 $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ at 23 ° C and 0 % RH.

(2) Eight apple wedges were packaged into polypropylene $< packaging >$ trays and wrap-sealed using a 64 $< numvalthick >$ μm $< um >$ thickness $< quantity >$ polypropylene $< packaging >$ film with a permeability to oxygen $< quantity >$ of 110 $< numvalperm >$ $cm^3\ m^{-2}\ bar^{-1}\ day^{-1}$ $< um >$ at 23 $< numvaltemp >$ ° C $< um >$ and 0 $< numvalrh >$ % $< um >$ RH $< quantity >$.

Now that we have proposed new textual contexts to explore and a new data representation that substantially increases argument expressiveness in text thanks to the OTR, we detail our knowledge discovery process to discover frequent patterns of correlated arguments.

3.3 Knowledge discovery process guided by a domain OTR

In this section, we propose a knowledge discovery process, illustrated in figure 3, which relies on four steps and is guided by a domain OTR.

Process description.

The first step corresponds to the data representation presented in section 3.2, and that allows to increase data expressiveness.

The second step proposes to represent several textual windows to be mined in the datamining step, using the units referenced in the OTR, and the definitions 3 and 4 we have previously proposed. Those textual windows draw specific textual contexts where relevant information about n-ary relations can be discovered.

Then, the third step prepares our textual windows for the datamining step. From the data representation proposed in section 3.2, we can define a transaction and item as follows :

Definition 11. *(Transaction)*
A transaction is defined as a set of sentences according to a textual window.

Example 4.
In a textual window $f_{\pm 1}$, each transaction corresponds to a set of sentences composed of the pivot sentence, the previous and afterward sentences.

Definition 12. *(Item)*
A set of items I_n is the set of the n nearest terms or concepts of the identified concepts in data representation.

Example 5.
Let us consider the sentence (2) of example 3, if we choose to select the 1-term nearest neighboors of the identified concept <packaging>, we obtain a set of items composed of <packaging>, polypropylene, trays, films.

Finally, the fourth step corresponds to the datamining step. Each represented textual window, according to the transactions and set of items, is transformed into a matrix to be explored by datamining algorithms in order to discover frequent patterns of correlated arguments as ontological sequential patterns (OSP) defined as follows:

Definition 13. *(Ontological Sequence)*
$\mathcal{W}$ is a set of words in a sentence;
$\mathcal{C}_{\mathcal{OTR}}$ is a set of concepts of the OTR;
$\mathcal{W}_{\mathcal{OTR}}$ is a set of words in the terminological component of the OTR where all words $w_i \in \mathcal{W}_{\mathcal{OTR}}$ denote a concept $c_i \in \mathcal{C}_{\mathcal{OTR}}$;
$\mathcal{IO} = \{IO_1, IO_2, ..., IO_m\}$ be a set of items IO_j where $IO_j \in \mathcal{W} \times \mathcal{W}_{\mathcal{OTR}}$ or $IO_j \in \mathcal{C}_{\mathcal{OTR}}$. An itemset is a non empty, non ordered set of items noted $(IO_1, IO_2, ..., IO_k)$ where $IO_j \in \mathcal{IO}$. An ontological sequence is a non empty ordered list of itemsets noted $< IOS_1 IOS_2 ... IOS_p >$ where $IOS_j \in \mathcal{IOS}$, with $\mathcal{IOS}$ the set of itemsets.

Extracting ontological sequential patterns consists in searching frequent ontological sub-sequences from ontological sequences.

Definition 14. *(Ontological Sub-sequence)*
An ontological sequence $A = < IOS_1 IOS_2 ... IOS_p >$ is an ontological sub-sequence of an other ontological sequence $B = < IOS'_1 IOS'_2 ... IOS'_m >$ $(A \preceq B)$ if $p \leq m$ and if there exists integers $j_1 < j_2 < ... < j_k < ... < j_p$ such as $IOS_1 \subseteq IOS_{j_1}, IOS_2 \subseteq IOS_{j_2}, ..., IOS_p \subseteq IOS_{j_p}$.

An Ontological Sequential Pattern (OSP) is a frequent ontological sub-sequence characterized by a support, which represents the number of occurrences of a pattern in $\mathcal{OS}$, a set of ontological sequences.

At the end of the knowledge discovery process, we automatically extract a set of ontological sequential rules and patterns of argument expression in text. Those patterns are based on the data representation that uses the conceptual level of the OTR and are generic patterns.

Example 6.
*Consider the OSP <(packaging)(numvalthick um)> supported by sequences $OS_{f_{\pm 1}}$ with $OS_{f_{\pm 1}}$ as $OS \preceq OS_{f_{\pm 1}}$. This OSP, discovered in the sequences of the textual window $f_{\pm 1}$, associates the **packaging concept** of food packaging domain OTR with the representation of **thickness value** and the **um concept** (unit concept).*

In the following, we propose a hybrid approach that takes advantages of syntactic relations to enrich the extracted OSP with relevant grammatical categories and terms (in our work a term refers to a word). As we will show in next section, relevant grammatical categories and terms are the ones that stay close to the expression of arguments involved in the searched relation.

4. USING SYNTACTIC RELATIONS

In this section, we present our hybrid approach that combines frequent OSP with syntactic analysis in order to construct linguistic sequential patterns of correlated arguments in text. We firstly present the basics of syntactic analysis and secondly explain our choices towards syntactic relations and finally detail our hybrid approach to construct correlated argument patterns.

4.1 Syntactic analysis basis

Syntactic analysis is the process of analysing a string in natural language conforming to the rules of a formal grammar. The analysis consists in segmenting a sentence into parts, e.g. noun phrase, verb phrase and into categories, e.g. noun, verb and in representing the phrasal structure using a tree representation where each word is represented according to its syntactic group or category. The syntactic analysis gives us frequent syntactic structures used and allows us to better understand how sentences are constructed in natural language. Despite progress in program development of syntactic analysis, there is not yet an efficient parser that covers the bulk of formal grammar, provides reasonable number of analysis per sentence and it is not limited to the length of sentences it analyses. Moreover, the phrasal structure often used in linguitic domain provides a limited semantic level of analysis. Thereby, in our work we didn't focus in first choice on syntactic analysis but rather on discovering implicit rules based on semantic analysis guided by the OTR. Then, in the hybrid approach, we are interested in improving the structure of our discovered patterns using specific syntactic relations, so that we are able to master the number and the quality of analysis in order to construct relevant patterns.

4.2 Syntactic relations

Our approach uses the advantages of datamining methods to discover implicit rules and patterns of correlated argu-

ments in text. As we have previously explained, those discovered patterns and rules only restitute the relationships that arguments share in text. Syntactic relations (SR) provide a simple description of the grammatical relationships in a sentence that can easily be understood and effectively used without linguistic expertise, especially in tasks involving information extraction from text. In particular, rather than the phrase structure representations that have long dominated in the computational linguistic community, it represents all sentence relationships as triples of a relation between pairs of words, such as in the sentence "Bell, based in Los Angeles, makes and distributes electronic, computer and building products". The subject of *distributes* is *Bell*. The representation chosen is the following SR: *nsubj(distributes, Bell)*.
In our work, we are interested in extracting all relevant SR that provide grammatical relationships between arguments involved in the discovered sequential patterns. To summarize, we can describe the hybrid approach, presented in next section, as a combination of data mining approaches, which provide exhaustive extractions of implicit relations and patterns of arguments in text, and specific SR to understand linguistic structures often used in text to describe arguments involved in those patterns.

4.3 Xart hybrid approach

In this section, our work focuses on extracting frequent and relevant SR close to the arguments of searched n-ary relations. Those SR reveal specific grammatical relations and specific terms used in the expressions of arguments in text. We propose to use those SR in order to construct linguistic sequential patterns to be applied on text to extract correlated arguments of n-ary relations.

The hybrid approach, as illustrated in figure 4 is composed of four main steps :

SR extraction.
consists in using a syntactic parser on a corpus. The parser analyses each sentence in the corpus and returns all SR. The SR links a gammatical category to a pair of words. For example, the RS *NN(thickness, film)* explains that the words *thickness* and *film* are linked with a syntactic nominal relation. In this work, our interest is extracting all SR that link words close to the arguments of n-ary relation. For that purpose, we use a domain OTR to identify the most relevant SR.

SR close to the OTR extraction.
We are interested in extracting all SR that are close to the OTR, i.e. the ones that contain at least one term denoting a concept in the OTR. We define those syntactic relations as Ontological Syntactic Relation (OSR).

Definition 15. *(OSR)*
$\mathcal{SR}$ is a set of syntactic relations;
$\mathcal{W}$ is a set of words in a sentence;
$\mathcal{W}_{\mathcal{OTR}}$ is a set of words in the terminological component of the OTR;
$\mathcal{C}_{\mathcal{OTR}}$ is a set of concepts of the OTR;

OSR is defined as a relation (w_1, sr, w_2) with $sr \in \mathcal{SR}$, where $w_1 \in \mathcal{W} \times \mathcal{W}_{\mathcal{OTR}}$ and $w_2 \in \mathcal{W} \times \mathcal{W}_{\mathcal{OTR}}$. All words $w_i \in \mathcal{W}_{\mathcal{OTR}}$ denote a concept $c_i \in \mathcal{C}_{\mathcal{OTR}}$ and $\mathcal{C}_{\mathcal{OTR}} \in \mathcal{C}_{core}^{down}$.

0SR close to OSP.
Then, from this set of OSR, we focus on those that link words expressing correlated arguments discovered in the OSP detailed in section 3. For example, prep_of(thickness, LDPE) explains that the words *thickness* and *LDPE* are linked with a syntactic prepositional relation, *prep_of*. *LDPE* is the name of a packaging and denotes the concept $< Packaging >$ in the OTR. This OSR can be used to enrich the discovered patterns of correlated *packaging* and *thickness* arguments. Those kinds of OSR are very interesting to use in order to obtain Ontological Linguistic Sequential Patterns (OLSP) of correlated arguments.

Construction of OLSP.
All 0SR with a pair of words that express correlated arguments are used in order to enrich sequential patterns. We define those enriched patterns as Ontological Linguistic Sequential Patterns (OLSP).

Definition 16. *(OLSP)*
$\mathcal{OSR}$ is a set of ontological syntactic relations;
$\mathcal{OSP}$ is a set of ontological sequential patterns;
$\mathcal{SR}$ is a set of syntactic relations;
$\mathcal{W}$ is a set of words in a sentence;
$\mathcal{C}_{\mathcal{OTR}}$ is a set of concepts of the OTR;
$\mathcal{W}_{\mathcal{OTR}}$ is a set of words in the terminological component of the OTR where all words $w_i \in \mathcal{W}_{\mathcal{OTR}}$ denote a concept $c_i \in \mathcal{C}_{\mathcal{OTR}}$;

For each OSR $\in \mathcal{OSR}$ defined as the relation (w_1, sr, w_2), with $sr \in \mathcal{SR}$, where $w_1 \in \mathcal{W} \times \mathcal{W}_{\mathcal{OTR}}$ and $w_2 \in \mathcal{W} \times \mathcal{W}_{\mathcal{OTR}}$,
For each OSP $\in \mathcal{OSP}$ defined as a frequent ontological subsequence from a set of items $\mathcal{IO} = \{IO_1, IO_2, ..., IO_m\}$ where IO_j can be w_1 or IO_j can be w_2,
An Ontological Linguistic Sequential Pattern(OSP) corresponds to the OSP enriched with the relation (w_1, sr, w_2).

For example, an OSP extracted from the corpus of packaging domain <(packaging)(numvalthick um)> shows that the expression of packaging and thickness arguments are correlated in text, and we try to figure out in which linguistic structure(s) they are correlated using OSR. From this pattern, we look for all OSR that provide the relations between the searched arguments in the pattern, e.g. *NN(thickness, film/films) or*
NN(film/films, thickness), NN(LDPE, film/films) or
NN(film/films, HPMC), prep_of(thickness, LDPE). LDPE and HPMC are terms that denote the concept $< packaging >$ in the OTR, we keep the concept in the relation. OSR are then transformed, using the rules of the formal grammar related to the used parser, in linguistic structures linking the words of the OSR. Linguistic structure generation rules are given in example 7.

Example 7.
Generation :
- NN(thickness, film/films) => film/films thickness

- NN(film/films, thickness) => thickness film/films
- NN(LDPE, film/films) => film/films (packaging)
- NN(film/films, HPMC) => (packaging) film/films
- prep_of(thickness, LDPE) => thickness of (packaging)
Combining with OSP
<(packaging)(numvalthick um)> :
- <(packaging) film/films thickness (numvalthick um)>
- <thickness of (packaging) film/films (numvalthick um)>
...

We obtain OLSP constructed according to the sequence defined in the OSP. Those OLSP are used on text in order to extract sentences where we now can find correlated arguments of n-ary relations, e.g. *mango <packaging> films thickness was 0.17 ± 0.02 <numvalthick> mm <um>* ou *Thickness of resulting starch <packaging> films ranged from 199.6 ± 22.6 to 271.4 ± 581 <numvalthick> μm < um >*. In this section, we have shown how the proposed hybrid approach used in Xart system takes advantages of syntactic analysis and uses the OTR for the extraction of OSR that are specific SR we can use to construct OSLP.

5. EVALUATION

We have led experimentations on a corpus in food packaging field. We have first applied our knowledge discovery process in order to discover ontological sequential rules and patterns of argument expressions, then, after a validation step, we have used selected syntactic relations in order to enrich OSP and obtain OLPS to extract correlated arguments of n-ary relations from text.

5.1 Ontological sequential patterns

Sub-corpus constitution. From the food packaging corpus, we organised several sub-corpora according to textual windows represented (e.g. a corpus f_0, $f_{\pm 2}$). We applied our knowledge discovery process and obtained several matrix for each sub-corpus tested. The number of transactions tested changes according to the textual window represented from 5 000 to 35 000. The number of items also changes according to the textual window represented from 2 000 to more than 10 000.

Algorithms used in experiments. A substantial amount of datamining algorithms exists to the state-of-the-art, such as Apriori [1], Spade [30] and PrefixSpan [22]. The experiments have been led using Clospan [29] to extract sequential patterns. Clospan implements the most efficient algorithm to the state-of-the-art, PrefixSpan, and allows to discover a set of sequential patterns without redondancy and without loss of informativeness. We use CMRules [12] to extract sequential rules.

Selection criteria. A well-known issue in datamaning is managing the number of sequential patterns and rules generated from the algorithms. Thus, the support is an important measure used to eliminate uninteresting sequential rules and patterns and can be exploited for the efficient discovery of sequential rules and patterns. Confidence is interesting to use in order to measure the reliability of a rule. In our experiments we have used the higher confidence because the higher the confidence, the more likely it is for a set of itemsets Y to be present in sequences that contain the set of itemsets X.

Beyond those classical measures of support and confidence, we propose to use two new selection criteria based on both statistical and semantic criteria. The first one will select only the OSP where at least one argument of n-ary relations represented in the domain OTR is identified. The second one will select the OSP from the intersection of several studied textual windows.

Quantitative results. The number of ontological sequential rules and patterns varies according to the selection criteria applied. For example, we obtained more than 52 000 rules and patterns on the sub-corpus $f_{\pm 2}$ according to a minimum support of 0.5 and the criteria of selecting the rules and patterns containing at least one argument referenced in the OTR. When we added the selection criteria of intersection, we reduced this number around 1 000 OSP and rules.

Qualitative results. We first applied the knowledge discovery process without increasing expressiveness of arguments with data representation in text. We obtained a small set of patterns comparing to other results, i.e. around 500, and extracted patterns were meaningless, e.g. none of the patterns restituted numerical values whereas they are important to discover new instances in text.

Textual window	Ontological sequential rule or pattern	Support
$f_{\pm 1}$	<(packaging)(numvalthick um)>	0.5
	<(numvalthick)(films)>	0.5
	<(film)(mm)(thickness)>	0.1
	<(film thickness)(rh)>	0.1
	<(packaging)(quantity)(permeability)>	0.5
	<(packaging)(permeability)>	0.6
f_0	<(pressure)(water permeability)>	0.05
	<(oxygen permeability)(pressure)>	0.05
	<(thickness)(films)(observed)>	0.05
	<(thickness)(water)(films)>	0.07
$\bigcap f_n$	packaging => numvalthick	>0.05
	temperature => numvalrh	
	<(numvaltemp)(numvalrh%)>	
	<(packaging)(numvalthick)>	
	<(packaging)(numvaltemp °c)>	
	<(packaging) => temperature numvalrh >	
	packaging => quantity numvaltemp °c	

Table 3: Excerpt of OSP and rules - $\bigcap$ window intersection criteria

Table 3 gives an excerpt of OSP and rules obtained with the knowledge discovery process using data representation. First, the results show the interest of the new data representation to extract more meaningful patterns and rules. Secondly, the results show that extracted rules and patterns allow us to discover implicit argument expressions in text. In the studied domain, we find the five following rules:

1. A first rule showed in several patterns, *<(packaging)(numvalthick um)>*, highlights that *packaging* and *thickness* arguments frequently appear as correlated in text and that correlation frequently occurs in a maximal textual window of $f_{\pm 1}$;

2. A second rule, *temperature => numvalrh*, highlights that if we find *temperature* argument, we frequently find *relative humidity* argument. Those correlated arguments in text frequently appear in the same sentence, i.e. f_0;

3. A third rule, *<(packaging) => temperature numvalrh >*, highlights that if we find *packaging* argument, we frequently find *temperature* and *relative humidity* arguments. Thus, four correlated arguments in text, i.e.

packaging, thickness, temperature and *relative humidity* frequently occur in a maximal textual window of $f_{\pm 1}$;

4. A fourth rule, <(pressure)(water permeability)>, shows that *partial pressure* argument and *permeability* argument frequently occur in the same sentence;

5. A last rule suggests that *packaging* argument could be the trigger word of the relation since it frequently occurs in rules and patterns of previous correlations, e.g. *<(packaging)(permeability)>*, *<(packaging)(numvaltemp °c)>*.

The discovered correlations are relevant in 1 000 OSP and rules in the studied domain. After being validated by a domain expert, OSP can be used to construct OLSP using Xart hybrid approach and can be used to extract instances of correlated arguments as detailed in the following.

5.2 From ontological sequential patterns to ontological linguistic sequential patterns

Quantitative and qualitative results on SR extraction. We first used Stanford parser [17], the most efficient tool to the state-of-the-art (with 86% of accuracy) to extract all SR from the corpus. We obtained more than 50 000 SR. Then we extracted the OSR, which represent a meaningful subset close to arguments involved in n-ary relations. Finally, we kept only the OSR containing words that suggest correlated arguments discovered in OSP. The OSR are added to the selected OSP in order to construct OLSP of correlated arguments in text. We reduced the subset to 6 600 meaningful OSR. In terms of qualitative results, the most meaningful OSR return syntactic nominal, prepositional, and conjunctional relations. Other OSR show that the relation between a numerical value and its unit is always according to a numerical type. Another subset of interesting OSR include syntactic verb relations that are verbs describing experimental context or experimental results such as *conduct, prepare, measure, pack, compare, increase, decrease.*

Quantitative results on correlated argument extraction. The evaluations have been led using both OSP and OLSP on a subset of 11 articles where 87 argument instances have been manually annotated. The results are shown in table 4. Those patterns were directly used on texts of the sub-corpus $f_{\pm 1}$ of food packaging corpus where we found four correlated arguments *packaging, thickness, temperature, and relative humidity (RH).* In this first evaluation, we only used six frequent OSP, given in remark 1, that were enriched with four frequent types of OSR *noun, numerical, prepositional, and conjunctional* groups to construct the OLSP. We finally had a set of 50 patterns to test on texts composed of OSP and OLSP constructed with Xart hybrid approach. The results given in table 4 are measured in terms of precision, recall, and F-measure. The results are given according to the evaluation type of correlated arguments. The results show that using Xart hybrid approach substantially increases the precision of extraction from 0.4 to 0.7 for the first ones and from 0.3 to 0.7 for the second ones without too much impact on the recall. The results on the extration of 3 or 4 correlated arguments are slightly better using OLSP but it clearly shows the strength of OSP to discover accurate correlations with a F-measure of 0.6.

Remark 1.
OSP used in the evaluations :
<(packaging)(numvalthick um)>
<(numvalthick um)(packaging)>
<(packaging)(numvaltemp °c)>
<(numvaltemp °c)(packaging)>
<(packaging)(numvalrh%)>
<(numvalrh%)(packaging)>

6. CONCLUSION

In this work, we have proposed Xart system based on a hybrid approach that takes firstly advantages of datamining techniques and secondly of syntactic analysis. We have proposed a knowledge discovery process that takes into consideration expressiveness of data using the conceptual level given by a domain Ontological and Terminological Resource (OTR). In this process, we also have proposed to evaluate several textual windows in order to analyse ontological sequential rules and patterns (OSP) in a precise sequence of sentences. In this first step, the ontological sequential rules and patterns have shown several correlated arguments in texts. Then, we have proposed to extract relevant syntactic relations called Ontological Syntactic Relations (OSR) that improved the effectiveness and expressiveness of ontological sequential patterns by combining different levels of word abstraction (word, syntactic relation and concept) in Ontological Linguistic Sequential Patterns (OLSP). Finally, we have conducted some experiments with OSP and OLSP to extract sentences where we find from two to four correlated arguments. Further work is to apply the complete Xart system on a new corpus using a new domain OTR but with issues that remain the same, i.e. extraction of relevant data modeled as n-ary relations and as defined in naRyQ OTR. Another further work is to integrate the OLSP in a tool, @web[1] software, that allows researchers to manually annotate tables extracted from published scientific documents [6]. Indeed, in the tables of articles that return experimental results on packagings, it often occurs that some arguments (e.g. thickness) are missing in the table and are given in the text of the article. Specific OLSP (e.g. packaging and thickness correlated arguments) can help researchers to complete the annotation of data given in tables.

7. ACKNOWLEDGMENTS

This work is partially funded by the **Labex NUMEV**, **INRA**, and the **3BCAR IC2ACV** project.

8. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[2] R. Agrawal and R. Srikant. Mining sequential patterns. In Proceedings of the Eleventh International Conference on Data Engineering, ICDE '95, pages 3–14, Washington, DC, USA, 1995. IEEE Computer Society.

[3] N. Béchet, P. Cellier, T. Charnois, and B. Crémilleux. Discovering linguistic patterns using sequence mining.

[1] http://www6.inra.fr/cati-icat-atweb/Web-platform

Evaluation type	OSP			OLSP		
	P	*R*	*F*	*P*	*R*	*F*
General evaluation	0.5	0.8	**0.6**	0.7	0.8	**0.6**
packaging and *thickness*	0.4	0.9	0.5	0.7	0.9	**0.8**
temperature and *relative humidity*	0.3	0.9	0.4	0.8	0.7	**0.7**
n > 2 correlated arguments	0.6	0.6	**0.6**	0.7	0.6	**0.6**

Table 4: OSP and OLSP assessments for correlated argument extraction - Precision (P), Recall(R), F-measure(F)

In Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I, CICLing'12, pages 154–165, Berlin, Heidelberg, 2012. Springer-Verlag.

[4] S. L. Berrahou, P. Buche, J. Dibie-Barthélemy, and M. Roche. How to extract unit of measure in scientific documents? In KDIR/KMIS 2013 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing, Vilamoura, Algarve, Portugal, 19 - 22 September, 2013, pages 249–256, 2013.

[5] J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski. Extracting complex biological events with rich graph-based feature sets. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, BioNLP '09, pages 10–18, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[6] P. Buche, S. Dervaux, J. Dibie-Barthélemy, L. Soler, L. Ibanescu, and R. Touhami. Intégration de données hétérogènes et imprécises guidée par une ressource termino-ontologique. Revue d'Intelligence Artificielle, 27(4-5):539–568, 2013.

[7] Q.-C. Bui and P. M. A. Sloot. Extracting biological events from text using simple syntactic patterns. In Proceedings of the BioNLP Shared Task 2011 Workshop, BioNLP Shared Task '11, pages 143–146, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[8] E. Buyko, E. Faessler, J. Wermter, and U. Hahn. Event extraction from trimmed dependency graphs. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, BioNLP '09, pages 19–27, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[9] P. Cellier, T. Charnois, M. Plantevit, C. Rigotti, B. Crémilleux, O. Gandrillon, J. Kléma, and J. Manguin. Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts. J. Biomedical Semantics, 6:27, 2015.

[10] L. Di-Jorio, S. Bringay, C. Fiot, A. Laurent, and M. Teisseire. Sequential patterns for maintaining ontologies over time. In On the Move to Meaningful Internet Systems: OTM 2008, OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Monterrey, Mexico, November 9-14, 2008, Proceedings, Part II, pages 1385–1403, 2008.

[11] M. Fabrègue, A. Braud, S. Bringay, F. L. Ber, and M. Teisseire. Including spatial relations and scales within sequential pattern extraction. In Discovery Science - 15th International Conference, DS 2012, Lyon, France, October 29-31, 2012. Proceedings, pages 209–223, 2012.

[12] P. Fournier-Viger, U. Faghihi, R. Nkambou, and E. M. Nguifo. Cmrules: Mining sequential rules common to several sequences. Knowl.-Based Syst., pages 63–76, 2012.

[13] Y. Hao, X. Zhu, M. Huang, and M. Li. Discovering patterns to extract proteinâĂŞprotein interactions from the literature: part ii. Bioinformatics, 21:3294âĂŞ3300, 2005.

[14] L. Hawizy, D. Jessop, N. Adams, and P. Murray-Rust. ChemicalTagger: a tool for semantic text-mining in chemistry. Journal of cheminformatics, 3(1):17, 2011.

[15] M. Huang, X. Zhu, D. G. Payan, K. Qu, and M. Li. Discovering patterns to extract protein-protein interactions from full biomedical texts. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications, JNLPBA '04, pages 22–28, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[16] S. Jaillet, A. Laurent, and M. Teisseire. Sequential patterns for text categorization. Intell. Data Anal., 10(3):199–214, 2006.

[17] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, pages 423–430, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[18] Q. Le Minh, S. N. Truong, and Q. H. Bao. A pattern approach for biomedical event annotation. In Proceedings of the BioNLP Shared Task 2011 Workshop, BioNLP Shared Task '11, pages 149–150, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[19] R. Mcdonald, F. Pereira, S. Kulick, S. Winters, Y. Jin, and P. White. Simple algorithms for complex relation extraction with applications to biomedical ie. In In Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL-05, pages 491–498, 2005.

[20] A.-L. Minard, A.-L. Ligozat, and B. Grau. Multi-class svm for relation extraction from clinical reports. In G. Angelova, K. Bontcheva, R. Mitkov, and N. Nicolov, editors, RANLP, pages 604–609. RANLP 2011 Organising Committee, 2011.

[21] M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii. A rich feature vector for protein-protein interaction extraction from multiple corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, EMNLP '09, pages 121–130, Stroudsburg, PA, USA, 2009.

Association for Computational Linguistics.

[22] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In Proceedings of the 17th International Conference on Data Engineering, pages 215–224, Washington, DC, USA, 2001. IEEE Computer Society.

[23] D. Proux, F. Rechenmann, and L. Julliard. A pragmatic information extraction strategy for gathering data on genetic interactions. In P. E. Bourne, M. Gribskov, R. B. Altman, N. Jensen, D. A. Hope, T. Lengauer, J. C. Mitchell, E. D. Scheeff, C. Smith, S. Strande, and H. Weissig, editors, ISMB, pages 279–285. AAAI, 2000.

[24] K. Raja, S. Subramani, and J. Natarajan. Ppinterfinder - a mining tool for extracting causal relations on human proteins from literature. Database, 2013, 2013.

[25] B. Rosario and M. A. Hearst. Classifying semantic relations in bioscience texts. In Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[26] B. Rosario and M. A. Hearst. Multi-way relation classification: Application to protein-protein interactions. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pages 732–739, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[27] R. Touhami, P. Buche, J. Dibie-Barthélemy, and L. Ibanescu. An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables. In OTM Conferences (2), pages 662–679, 2011.

[28] S. Van Landeghem, Y. Saeys, B. De Baets, and Y. Van de Peer. Analyzing text in search of bio-molecular events: A high-precision machine learning framework. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, BioNLP '09, pages 128–136, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[29] X. Yan, J. Han, and R. Afshar. Clospan: Mining closed sequential patterns in large databases. In D. Barbará and C. Kamath, editors, SDM. SIAM, 2003.

[30] M. J. Zaki. Spade: An efficient algorithm for mining frequent sequences. Mach. Learn., 42(1-2):31–60, Jan. 2001.

[31] H. Zhang, M. Huang, and X. Zhu. Protein-protein interaction extraction from bio-literature with compact features and data sampling strategy. In 4th International Conference on Biomedical Engineering and Informatics, BMEI 2011, Shanghai, China, October 15-17, 2011, pages 1767–1771, 2011.

[32] D. Zhou, D. Zhong, and Y. He. Biomedical relation extraction: From binary to complex. Comp. Math. Methods in Medicine, 2014, 2014.